B.6. Búsqueda semántica en *Europeana*: se percibe el problema pero aún no la solución

Por **Ernest Abadal** y **Lluís Codina**

29 enero 2009

Abadal, Ernest; **Codina, Lluís**. "Búsqueda semántica en *Europeana*: se percibe el problema pero aún no la solución". *Anuario ThinkEPI 2009*, EPI SCP, pp. 55-58.

Resumen: *Internet ha propiciado la creación de grandes colecciones de documentos por medio de la agregación de fondos procedentes de diversos proveedores. Europeana es un ejemplo de ello. En general estos fondos han sido catalogados utilizando distintos lenguajes documentales y a la vez distintos idiomas. No existen aún sistemas que faciliten al usuario la consulta de estos documentos utilizando el lenguaje natural y en su propio idioma. Europeana ha puesto en funcionamiento un prototipo de motor de búsqueda semántico del cual se analiza su estructura y funcionamiento. La conclusión no obstante es que nos encontramos ante una versión aún muy experimental que no da satisfacción a las mínimas demandas de los usuarios.*

Palabras clave: *Europeana, Búsqueda semántica, Acceso temático multilingüe*.

Title: *Semantic search in* Europeana*: problem perceived but not the solution*

Abstract: *The internet has facilitated the creation of large collections of documents by the aggregation of documents sent from different suppliers. Europeana is an example. In general, these documents have been identified using different thesauri and, at the same time, different languages. There are no systems that help the user to consult these documents using natural language, and also in their own language. Europeana has set up a prototype of a semantic search engine. In this text we analyze its structure and functioning. The conclusion, however, is that this is a very experimental version.*

Keywords *Europeana, Semantic search, Multilingual subject access.*

Introducción

Con la buena noticia de la reapertura de *Europeana*, además de una búsqueda simple y avanzada disponemos de una aplicación experimental que quiere explorar vías para facilitar el acceso temático multilingüe a las colecciones. Esta (relativamente) nueva forma de búsqueda se denomina "*Europeana's semantic search engine*" y está aún en fase beta. Se puede acceder a ella desde un enlace directo o a través de la opción "Laboratorio de ideas" (*Thought Lab*) de la página principal de *Europeana*.

http://eculture.cs.vu.nl/europeana/session/search
http://www.europeana.eu/portal/

Con esta acción se puede comprobar que los responsables de *Europeana* han detectado a la perfección cuál es el problema aunque, como se verá, se encuentran aún lejos de encontrar la solución.

El problema: el acceso temático a una colección muy diversa

Como es bien sabido, los fondos de *Europeana* provienen de distintas bibliotecas, museos y colecciones audiovisuales europeas. En total actualmente contiene 2 millones de "ítems digitales", aunque se calcula que llegarán a ser más de 10 millones a lo largo de los próximos años[1]. Ahora bien, los registros de cada uno de los documentos, aunque puedan compartir el mismo esquema de metadatos (*Dublin Core* u otros), utilizan lenguajes documentales distintos para la representación del contenido y además están escritos en idiomas distintos.

En estos momentos ya resulta difícil la búsqueda debido a la disparidad de fuentes, idiomas, tipos de objetos digitales, etc. En el futuro, a pleno rendimiento de la colección, ¿cómo se podrá facilitar el acceso al usuario en estas condiciones?

¿Cómo resolver el problema derivado de la diversidad lingüística y documental?

La red facilita y tiende a la integración de colecciones. Lo más habitual es que cada colección tenga un sentido y entidad por sí misma, utilice un determinado idioma y quizás un determinado lenguaje documental. Cuando las agrupamos sin más y las ponemos a disposición del usuario, éste no puede aprovechar el potencial que le podrían ofrecer los metadatos que cada objeto o documento contiene.

"En el futuro, a pleno rendimiento de la colección, ¿cómo resolver el problema derivado de la diversidad lingüística y documental?"

El usuario que consulta una base de datos no tiene suficiente con disponer de los menús y opciones en su idioma (cosa que ya es posible actualmente en *Europeana*). El usuario necesita además poder utilizar su propia lengua cuando introduce los términos de búsqueda. ¿Qué sucede cuando buscamos información sobre "coches" en *GoogleBooks* si queremos obtener documentos también en catalán, francés o inglés?, pues que tenemos que utilizar el término en diversos idiomas (*car*, *cotxe*, *voiture*, etc.). Si además se usaran diversos lenguajes documentales (*subject headings*, descriptores, etc.) entonces tendríamos que saber si el término preferente (descriptor) es "automóvil", "turismo" o "vehículo" (y esto ¡para cada uno de los idiomas!).

¿Cómo solventar este doble problema? La resolución puede llevarse a cabo fundamentalmente de dos maneras: la traducción automática de las consultas (*cross-language text retrieval*) o mediante la utilización de lenguajes documentales (*multilingual subject access*).

La primera de estas vías (*cross-language text retrieval* o *multilingual text retrieval*) se investiga fundamentalmente desde el ámbito de la informática y se basa en el desarrollo de sistemas automáticos de traducción de las consultas o de expansión semántica hacia términos en otros idiomas. **Oard** y **Dorr** (1996) y **Oard** (1997) han elaborado un amplio y detallado estado de la cuestión de todos los estudios e investigaciones llevados a cabo bajo esta orientación.

En el caso de *Europeana*, parece que se opta por la segunda orientación. Es un modelo bastante diferente al anterior ya que no hay traducción automática de los términos de consulta, sino que se trata de ver la concordancia entre los términos introducidos por el usuario y los que forman parte del lenguaje documental multilingüe de que dispone el catálogo o base de datos y, más concretamente, desarrollando sistemas de equivalencias (mapeo).

El establecimiento de equivalencias, interoperabilidad o mapeo se refiere a la posibilidad de consultar de manera simultánea diversos fondos que han sido indexados con lenguajes documentales diferentes (el idioma puede ser una de estas diferencias, pero no es la única). En este caso, el énfasis se pone en desarrollar sistemas de equivalencias (mapeo) entre los términos de diferentes lenguajes documentales ya existentes.

La interoperabilidad (mapeo) entre diferentes lenguajes documentales implica establecer equivalencias entre términos de lenguajes documentales de estructura, lengua o grados de profundización diferentes. Esto explica que actualmente se encuentren pocas experiencias en funcionamiento de este modelo. Una primera experiencia fue *Macs* (*Multilingual Access to Subjects*), un prototipo creado en 1997 como respuesta a un encargo de la *Conference of European National Libraries* (*Cenl*) con el objetivo de encontrar soluciones al acceso multilingüe por materias a bases de datos bibliográficas. El proyecto *Macs* pretendía proporcionar acceso multilingüe por materias (en inglés, francés y alemán) a diferentes catálogos simultáneamente: el catálogo de las bibliotecas nacionales de Suiza, Francia, Gran Bretaña y Alemania.

Macs parte del convencimiento de que es posible navegar a través de vocabularios gracias a un sistema de enlaces de equivalencia entre los tres lenguajes de indexación (listas de encabezamientos) utilizados en las bibliotecas implicadas en el proyecto: *SWD* (Alemania), *Rameau* (Francia) y *Lcsh* (Inglaterra). *https://macs.vub.ac.be/pub/*

El prototipo

El prototipo que presenta *Europeana* no utiliza de hecho los fondos de *Europeana,* sino una colección de unos 150.000 registros de obras de arte de tres museos: *Rijksmuseum Amsterdam*, *Louvre* y el *Instituto de Historia del Arte de los Países Bajos*. Además se utilizan diversos tesaurus (*Joconde*, *IconClass*, *AAT*, *RKDArtists*, *WordNet*) que suman más de un millón y medio de términos (referidos tanto a conceptos como a personas, localizaciones y eventos).

La petición de un término permite lanzar búsqueda multilingüe y, además, de términos relacionados. Así por ejemplo, si buscamos "Paris"" nos presenta los documentos en los que aparece esta palabra (literalmente) ya sea en el título o en otras partes del registro y también aquellos otros

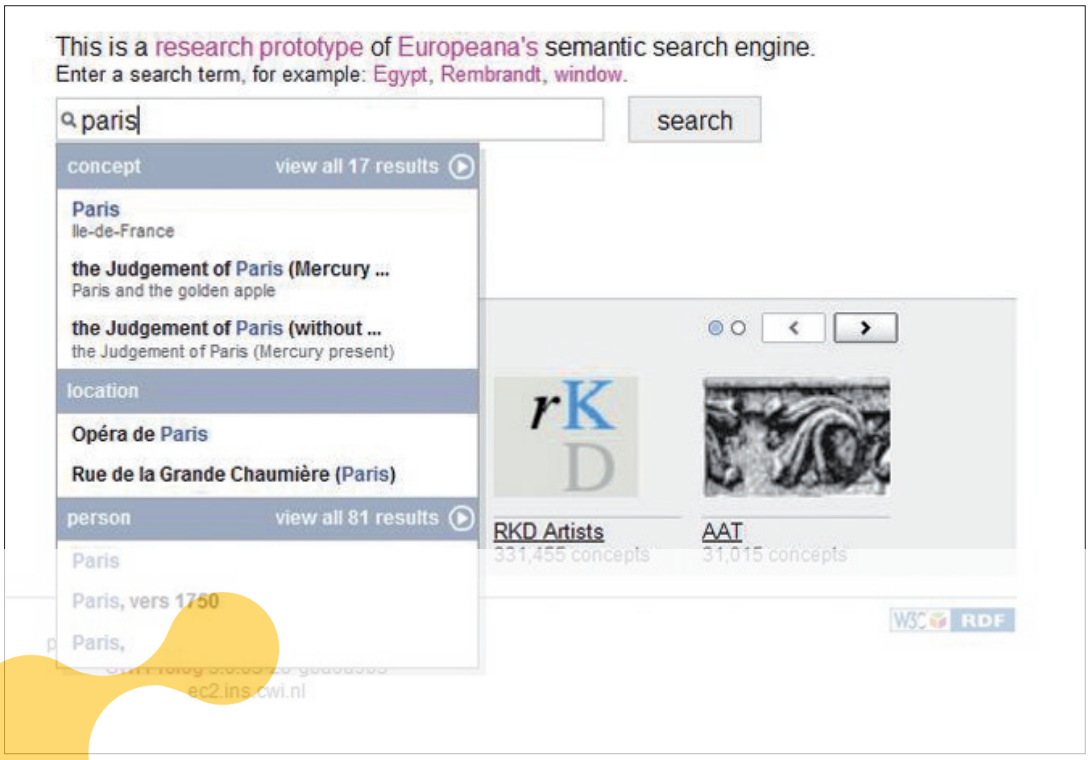

Figura 1. La desambiguación del término en la búsqueda por la palabra clave "Paris"

Figura 2: Otro ejemplo de desambiguación: en este caso por el término "David". Una de las acepciones más relevantes (el pintor Jacques-Louis David, sin embargo, no aparece)

en los cuáles se ha realizado una expansión semántica.

El proceso en concreto es el siguiente:

1. Introducción del término de búsqueda.

2. El sistema responde ofreciendo una relación de términos (de tesauros, títulos, etc.) en los cuales aparece el término de búsqueda organizado en diversas categorías, por ejemplo, si el término es "Paris", aparecerá "Paris" como lugar, como nombre propio, etc.

3. Una vez el usuario selecciona una de las opciones (p.e. "Paris" como lugar), se ejecuta la búsqueda.

4a. La página de resultados agrupa los ítems por diversas categorías o conceptos. En el caso de Paris (en el sentido de lugar) muestra decenas de categorías como las siguientes:

– Obras mostrando cosas más específicas de (Paris).

– Obras creados por personas que fallecieron en (Paris).

– Obras relacionados con personas que fallecieron en (Paris).

– Etc.

El problema es que resulta difícil interpretar algunas de las categorías presentadas. Por ejemplo: "obras creadas por un estudiante de una persona que falleció en (Paris)"; y aún las hay más retorcidas, como "obras relacionadas con un artista profesionalmente relacionado con la persona que falleció en (Paris)".

4b. En cambio, si buscamos por un término como "*woman*", los resultados incluyen agrupaciones como:

– Obras que muestran el concepto en el título.

– Obras que muestran un concepto más específico.

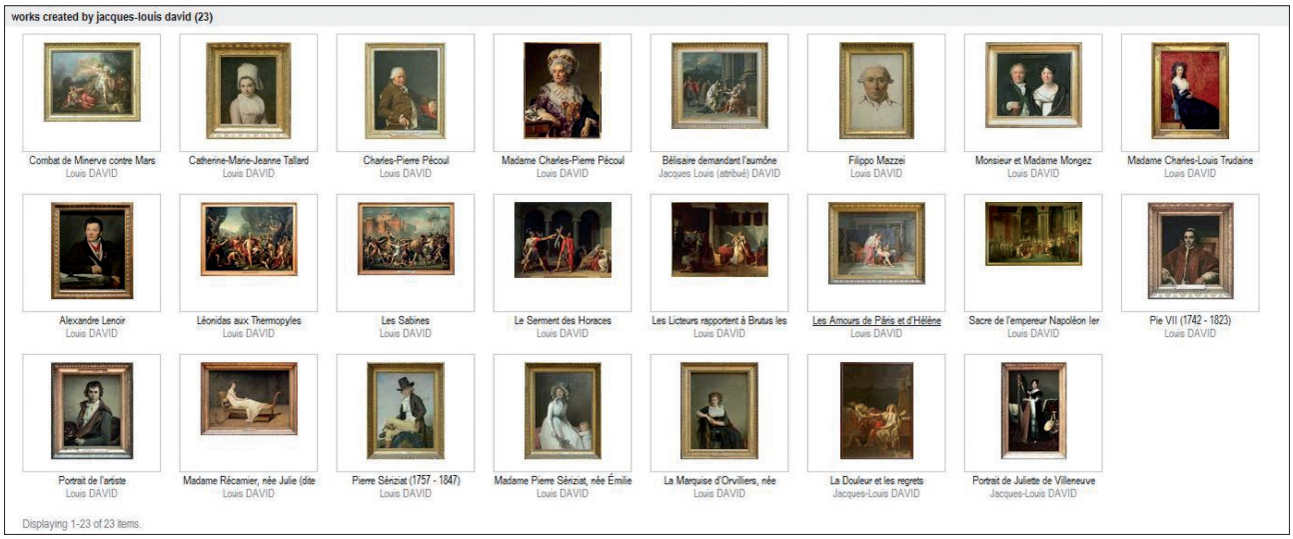

Figura 3: Un ejemplo de página de resultados (para el término Jacques-Louis David)

– Obras relacionadas con un artefacto que muestra el concepto.

De nuevo obtenemos una lista de categorías que progresivamente van mostrando un carácter menos comprensible.

Al parecer el algoritmo que combina los criterios que permiten elaborar esta clasificación de modo automático está todavía lejos de dar resultados válidos para usuarios finales. De hecho, para cualquier clase de usuario.

El programa utilizado, que se encuentra aún en fase inicial de desarrollo, se llama *ClioPatra* y dispone de diversas instalaciones experimentales. *http://e-culture.multimedian.nl/software/ClioPatria.shtml*

Es un desarrollo de *MultimediaN*, una organización sin afán de lucro radicada en los Países Bajos, con socios públicos (universidades) y otros privados. Tiene diversos proyectos en marcha relacionados con la búsqueda en internet, uno de los cuales es *N9 Eculture project*. En él participan informáticos de diversas universidades holandesas. *http://www.multimedian.nl*

Conclusiones

Como se puede comprobar, nos encontramos ante una propuesta poco madura. Algunos pueden considerar que no es ni siquiera una versión beta, sino una versión alfa o incluso de laboratorio y que quizá se ha pecado de una cierta precipitación, como ya se demostró con la ineficaz plataforma elegida para su lanzamiento. Por otro lado, no se puede dejar de indicar que todo el sistema de búsqueda (incluida la consulta avanzada y la presentación de los resultados) adolece de notables problemas de usabilidad.

"El proyecto *Europeana* tiene que continuar creciendo y ampliándose tanto en lo que se refiere a contenidos como también a las prestaciones de búsqueda"

En cualquier caso, se tiene que destacar que los responsables de *Europeana* demuestran que [illegible] y, aunque da claro que estamos aún lejos de encontrar una buena solución. El proyecto *Europeana* es magnífico y tiene que continuar creciendo y ampliándose tanto en lo que se refiere a contenidos como también a las prestaciones de búsqueda. Esperamos que en el futuro podamos disfrutar de estas mejoras.

Notas

1. La lista de las instituciones colaboradoras de *Europeana* puede consultarse en: *http://www.europeana.eu/portal/partners.html*

Bibliografía

Oard, Douglas W. "Cross-language information retrieval bibliography", 1997. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.1982*

Oard, Douglas W.; **Dorr, Bonnie J.** "A survey of multilingual text retrieval", April 1996. *http://www.lib.umd.edu/drum/handle/1903/807*